Beyond the dictionary: creating a long-term corpus resource

Deryle Lonsdale
Brigham Young University
lonz@byu.edu

The dictionary

- Learner's frequency dictionary for French
- Part of series including other languages (Spanish, Portuguese, German, Mandarin)
- Top 5,000 word families with associated information
- Main listing is by frequency
- Other indexes: alphabetical, POS
- Thematic vocabulary lists

The corpus

- Collected from various sources
- 23 million words: half spoken, half written
- Recent usage: post-1950
- No deliberate demographic proportioning
- Various genres but not entirely balanced

The problem

- Large quantity of data, almost 11,000 files
- One person did all data manipulation, programming, management
- Wide variety of computational tools, techniques
  - Perl, awk, sed, grep, make
- Want to create a corpus that can be used for linguistic research (and by others)

What next?

- Migrating to relational database architecture
- Correcting annotations
- Analysis beyond single words
- Packaging infrastructure for continued updating over time
- Web deployment

Conversion to db format

- Defining database schema for content
- Loading data into database (mySQL)
- Query via mySQL statements (for now)
  - Embed in more standard GUI?

Refining tokenization

- Linguistic issues:
  - Accents on uppercase letters
  - Word segmentation (dis-moi vs. week-end, l'homme vs. aujourd'hui)
- Extralinguistic issues
  - Punctuation
  - Symbols
  - Markup

Refining POS tags

- Used state-of-the-art hybrid approach integrating results from various taggers
- Editing/hand-correcting
  - Of questionable value for the dictionary
  - Indispensable for a linguistic corpus
- Occasional morphological errors
  - Allez, levez-vous
    VER:pres aller 0.792818 VER:impe aller 0.207182

Refining lemmatization

- Morphological ambiguity
- Some cases are tricky
  - Non-finite forms (p.part.: verb? gerund? adj?)
  - Abbreviations, case folding, symbols (&, %, Xe)
- Occasional homonymy
  - je suis, avec exactement le sens posé par Greenberg
    VER:pres suivre|être 0.999980

    Je suis une femme seule.
    VER:pres suivre|être 0.999980

The CCASH framework

- Cost-conscious Annotation Supervised by Humans
- Interactive annotation engine
- Java based, flexible, multilingual
- Discovers most difficult/surprising annotations
  - Maximum entropy Markov model (MEMM)
- User interacts to correct/verify, system learns and redoes computation/discovery

Genre analysis

- Text only: aéronautique, bouleversement, coïncider, crépuscule, guérilla, itinéraire, jadis, laïque, logistique, météorologique, microphone, solennel
- Spoken only: abusif, allô, bah, cafard, cingler, clown, copine, dingue, flic, fric, hockey, lucratif, machin, météo, micro, ouais, porno, sexy, sympa

Dialects: Canadian terms (1)

- Words only found in Canadian-related sources of corpus
  - Many items just due to data sparsity
  - Still, interesting items to consider…
- Local flora/fauna: béluga, capelan, couguar, dicotylédone, gadidé, huard, narval, outarde, wapiti
- Occupations: andrologie, céréaliculture, constable, économétricien, fruiticulteur, garde-pêche
- Food: arrow-root

Dialects: Canadian terms (2)

- Social trends: biculturalisme, constitutionnal-isation
- Pejoratives: béni-oui-oui, lèche-bottes, tramp
- Medical treatments: botulinique, détoxiquer
- Telecommunications: câblodistribution, télé-enseignement
- Geography: circumpolaire, terre-neuvien, transcanadien, westerner
- Politics: antiparlementaire, créditiste, législativement, néo-canadien, parlementairement, partisanerie, politicaillerie, vice-royal, whip

Dialects: Canadian terms (3)

- Ships/buildings: crabier, igloo, pipelinier
- Adjectives: épeurant, tannant
- Products: alcooltest, porte-bébé
- Fun words: cafouillis, entourloupette, vasouillard, vexateur
- Loanwords: dépendamment, slang, star-system, understatement

Collocations, MWE's

il y a
projet de loi
n'est pas
est-ce que
qu'est-ce
n'a pas
il s'agit
milliards de dollars
tout le monde
il y avait
je ne sais
ce qui concerne
je pense que
le gouvernement fédéral
n'ont pas
des nations unies

Preliminary levels of parsing

- Named entity recognition
- Verb frames
- Shallow parsing
- Analysis of native speaker errors (especially gender)

Conclusions

- Upgrading lexical resource to linguistic resource
- Extending beyond one-time application
- Use of state-of-the-art annotation tools
- Compelling work at the intersection of humanities and sciences

Questions?